Trustworthy AI

Richa Singh
IIT Jodhpur, India
richa@iitj.ac.in

Mayank Vatsa
IIT Jodhpur, India
mvatsa@iitj.ac.in

Nalini Ratha
University of Buffalo, USA
nratha@buffalo.edu

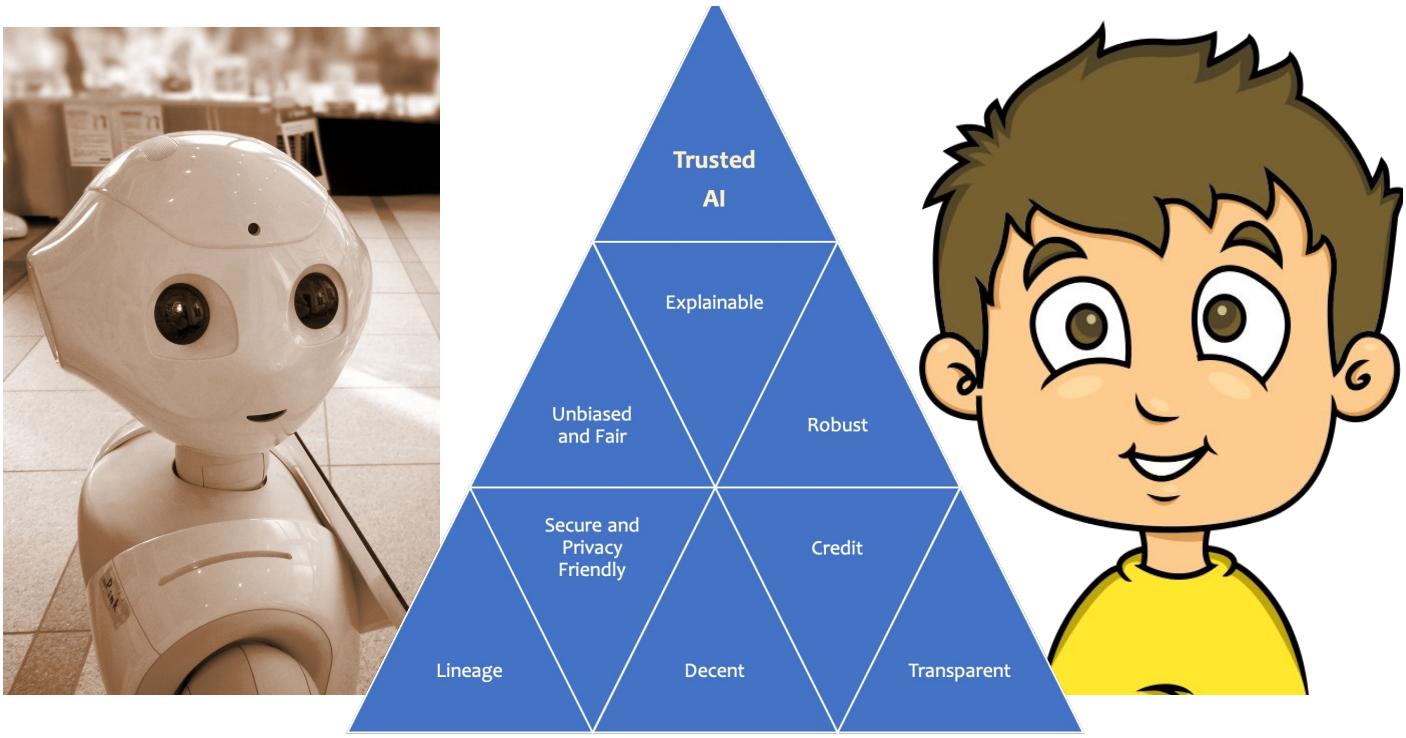

Figure 1: Components of AI Systems being Trustworthy and Decent.

ABSTRACT

Modern AI systems are reaping the advantage of novel learning methods. With their increasing usage, we are realizing the limitations and shortfalls of these systems. Brittleness to minor adversarial changes in the input data, ability to explain the decisions, address the bias in their training data, high opacity in terms of revealing the lineage of the system, how they were trained and tested, and under which parameters and conditions they can reliably guarantee a certain level of performance, are some of the most prominent limitations. Ensuring the privacy and security of the data, assigning appropriate credits to data sources, and delivering decent outputs are also required features of an AI system. We propose the tutorial on "Trustworthy AI" to address six critical issues in enhancing user and public trust in AI systems, namely: (i) bias and fairness, (ii) explainability, (iii) robust mitigation of adversarial attacks, (iv) improved privacy and security in model building, (v) being decent, and (vi) model attribution, including the right level of credit assignment to the data sources, model architectures, and transparency in lineage.

CODS COMAD 2021, January 2–4, 2021, Bangalore, India

ACM ISBN 978-1-4503-8817-7/21/01…$15.00
https://doi.org/10.1145/3430984.3431966

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Security and privacy**;

KEYWORDS

bias and fairness, explainability and interpretability, robustness, privacy and security, decent, transparency

ACM Reference Format:
Richa Singh, Mayank Vatsa, and Nalini Ratha. 2021. Trustworthy AI. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3430984.3431966

1 INTRODUCTION

Artificial intelligence (AI) systems are becoming an integral part of our daily lives. They are being employed for making mundane day to day decisions such as healthy food choices and dress recommendation, as well as for important and impactful decisions such as diagnosis of diseases, detection of financial frauds, and selecting new employees. Their increasing deployment in upcoming applications such as autonomous driving, automated financial loan approval, and cancer treatment recommendations have many worrying about the level of trust associated with AI today. Such concerns are genuine as many weaker sides of modern AI systems have been exposed through adversarial attacks, bias, and lack of explainability in the current rapidly evolving AI systems. Therefore, it is important to build mechanisms and approaches for "Trustworthy AI" systems.

Building a trustworthy AI system requires understanding if the model is biased or not. Bias has been a critical Achilles' heel problem for modern AI systems. Many applications from face recognition to language translation have shown high levels of bias in the systems, as demonstrated in non-uniform performance across different groups and test sets. This has strong implications on the fairness and accountability of such systems, with extremely significant societal implications. Explainability and interpretability are a requirement for such systems in many different contexts, for example law enforcement and medical, where black box decision making is not acceptable. Even though modern AI systems have reported high levels of accuracy, they are unable to explain their decision process and the causes of failures or successful cases to humans. Privacy and security are critical to success of AI beyond high accuracies. Recent research has shown that AI algorithms can exploit information extracted from social media to de-anonymize blurred faces, and promote unwanted spying through surveillance cameras. Such AI applications present both challenges and opportunities: while surveillance systems enhance safety of individuals and society at large, their susceptibility to attacks and breaches also provide the opportunity for abuse. Adversarial attacks in particular have created a huge negative perception with users that AI systems can be fooled easily. As researchers, we need to establish and promote a rigorous framework to formulate the problems in adversarial machine learning, evaluate the impact and consequences under various adversarial attacks, and characterize the properties that ensure the security of AI models.

As it has been observed in many areas, openness helps in unlocking larger potentials. Many of the AI systems do not disclose the lineage of the model, training data and performance details. More research is needed to address a common minimum acceptable practice to be disclosed by the systems. Data and model attribution are critical components of trusting an AI system. Accurate descriptions of training data, architecture and reliable test conditions are critical to guarantee a level of performance within predefined ranges, set users expectations and potentially explain potential biases and failures. Furthermore, especially in complex AI systems made of multiple components, attribution of a given prediction or signal from one specific model is fundamental for explainability, as well as security. Being able to reliably recognize the signature of an AI system provides a robust method to identify tampering, errors and potential breaches.

Furthermore, AI systems have barely benefited from recent advances in security such as fully homomorphic encryption. Recently, blockchain technology has been revolutionizing the way transactions are conceived, executed, managed, and monetized. While the commercial benefits of blockchain infrastructure are imminent, the applicability and advantages that the technology can offer to AI researchers and systems is still to be explored. As the world moves towards increasing decentralization of AI and emergence of AI marketplaces, a blockchain-based infrastructure would be essential to create the necessary trust between diverse stakeholders.

Another important aspect of being a human like AI system is being "decent"; though it has not being discussed in the literature directly. While designing an AI system, it is expected that the social behaviors that are habitual in human to human interactions are followed in AI agent-human interactions. We generally train on a noisy corpus and do not train/enforce it to be decent, which may lead to indecent responses by the agent. Several instances in the past have shown indecent behavior by the AI systems [12, 17, 26] and many of such systems have been decommissioned due to their indecent responses. Therefore, we believe that in order to build a trustworthy AI system, it is very important to have ensure that the system is *decent*.

Although researchers are studying individual challenges of AI systems independently, we believe that we need a comprehensive attempt to bring together all the aspects of trustworthy in AI under one umbrella and understand the interlinkages between them. This tutorial will discuss the fundamentals of trustworthy and decent AI, and summarize key and recent contributions in this area as well as open challenges. As shown in in Figure 1, we focus on the following topics:

- Bias and Fairness (estimation and mitigation)
- Adversarial Robustness (detection and mitigation)
- Explainability and Interpretability
- Trustworthiness and Security via Blockchain
- Privacy Preserving AI
- Credit, Lineage, and Transparency Approaches
- Decent AI

2 BUILDING TRUSTED/TRUSTWORTHY AI SYSTEMS

The EU guidelines on defining Trustworthy AI lays out seven requirements [1]:

(1) "*Human agency and oversight, including fundamental rights, human agency and human oversight*".
(2) "*Technical robustness and safety, including resilience to attack and security, fallback plan and general safety, accuracy, reliability and reproducibility*".
(3) "*Privacy and data governance, including respect for privacy, quality and integrity of data, and access to data*".
(4) "*Transparency, traceability, explainability and communication*".
(5) "*Diversity, non-discrimination and fairness, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation*".

(6) "*Societal and environmental well-being, sustainability and environmental friendliness, social impact, society and democracy*".
(7) "*Accountability, auditability, minimisation and reporting of negative impact, trade-offs and redress*".

There are several directions that researchers have pursued for addressing the above mentioned components. For instance, bias and fairness research has focused on understanding and estimating bias as well as mitigating and accounting for bias [10, 11, 16, 34–37]. It has been discussed in great detail in several papers [18, 32, 36]. Similarly, research on robustness against adversarial attacks has focused on generating attack models, detecting adversarial perturbations, and mitigating these attacks [6–8, 22–24, 28–31, 40, 42, 45, 47, 48]. Recently, different surveys have also highlighted the research progress in adversarial robustness [39, 43, 49]. In recent research threads, deep learning models are protected via cryptography measures to convert the internal layers of CNN into blocks of the blockchain [19–21]. Researchers are also working towards providing certified defense against adversarial perturbations. However, most of them are evaluated on the first-order adversary or gray-scale images [38, 46]. Cohen et al. [15] have shown the certified robustness of large scale ImageNet images and proved slight robustness for l_2 attacks. Moreover, research suggests that adversarial training is vulnerable to black-box attacks with several privacy issues and creates blind-spots for further attacks [33, 50]. In other research directions, explainability [9, 14, 27], transparency [25], data lineage [51], privacy and security [13], and social well being [41] have also been addressed.

2.1 Comparison with Similar Concepts

- Robust AI: In computer science, robustness is defined as the "*ability of a computer system to cope with errors during execution and cope with erroneous input*" [5].
- Ethical AI: The ethics of artificial intelligence, as defined in [3], "*is the part of the ethics of technology specific to robots and other artificially intelligent entities. It can be divided into a concern with the moral behavior of humans as they design, construct, use and treat artificially intelligent beings, and machine ethics, which is concerned with the moral behavior of artificial moral agents (AMAs). It also includes the issues of singularity and superintelligence.*" An AI system that follows this ethics guidelines would be considered as an Ethical AI system.
- Fair (unbiased) AI: "*In machine learning, a given algorithm is said to be fair, or to have fairness, if its results are independent of the given variables, especially those considered sensitive, such as the traits of individuals which should not correlate with the outcome (i.e. gender, ethnicity, sexual orientation, disability, etc)*" [4]. We would call an AI system fair or unbiased if it meets this criterion.
- Safe AI: Safe AI can be defined as a system that supports AI safety [2] - "*Artificial Intelligence (AI) Safety can be broadly defined as the endeavour to ensure that AI is deployed in ways that do not harm humanity*".
- Dependable AI: Dependable AI is defined as a system which is reliable, verifiable, explainable, and secure. This connects above mentioned themes and expands on the reliability and verifiability.

3 DECENT AI

We postulate that a truly intelligent system should be decent. In this context, we try to explore how can machines be trained to exhibit decent behavior. In order to define how to evaluate an AI system to be decent or indecent, we present the concept of Decent AI Turing Test.

Decent AI Turing Test: Inspired from the Turing test, as shown in Figure 2, we extend the original Turing test to Decent AI Turing Test. A human evaluator poses a question to the system and s/he does not know if the answer is given by a decent human agent or an automated AI agent. The response is then judged by the evaluator if it is correct or not and if it is decent or not. If the evaluator is not able to reliably differentiate the AI agent's response with the decent human agent, the AI agent is said to have passed the Decent AI Turing Test. More formally, we define:

A Decent AI Turing Test is a method of inquiry in AI for determining whether or not a computer is capable of thinking and behaving like a decent human being.

With this definition, we can argue that Decent AI is conceptually very different and complements all the characteristics of trustworthy AI systems. By that we mean that a decent AI system can have the properties of being fair, dependable, and trusted additionally.

4 RESEARCH QUESTIONS

There are a series of open research questions that can be explored. For instance:

(1) How to automatically measure/evaluate trustworthiness and decency of an AI system?
(2) How to integrate trustworthiness and decency with the performance/accuracy of the system?
(3) How to analyze the relationship between decency and different factors of trustworthy and dependable AI [44]?
(4) How to democratize trustworthy and decent AI systems?

5 BRIEF BIOGRAPHIES

Richa Singh received the M.S. and Ph.D. degree in computer science from West Virginia University, Morgantown, USA. She is currently a Professor at IIT Jodhpur, India. She is a Fellow of IAPR and a Senior Member of IEEE and ACM. She was a recipient of the Kusum and Mohandas Pai Faculty Research Fellowship at the IIIT-Delhi, the FAST Award by the Department of Science and Technology, India, and several best paper and best poster awards in international conferences. She is/was the Program Co-Chair of IJCB2020, FG2019 and BTAS 2016, and a General Co-Chair of FG2021 and ISBA 2017. She is also the Vice President (Publications) of the IEEE Biometrics Council and an Associate Editor-in-Chief of Pattern Recognition.

Mayank Vatsa received the M.S. and Ph.D. degrees in Computer Science from West Virginia University, USA. He is currently a Professor with IIT Jodhpur, India, and the Project Director of the Technology and Innovation Hub on Computer Vision and Augmented & Virtual Reality under the National Mission on Cyber Physical Systems by the Government of India. He is the recipient of the prestigious Swarnajayanti Fellowship from the Government

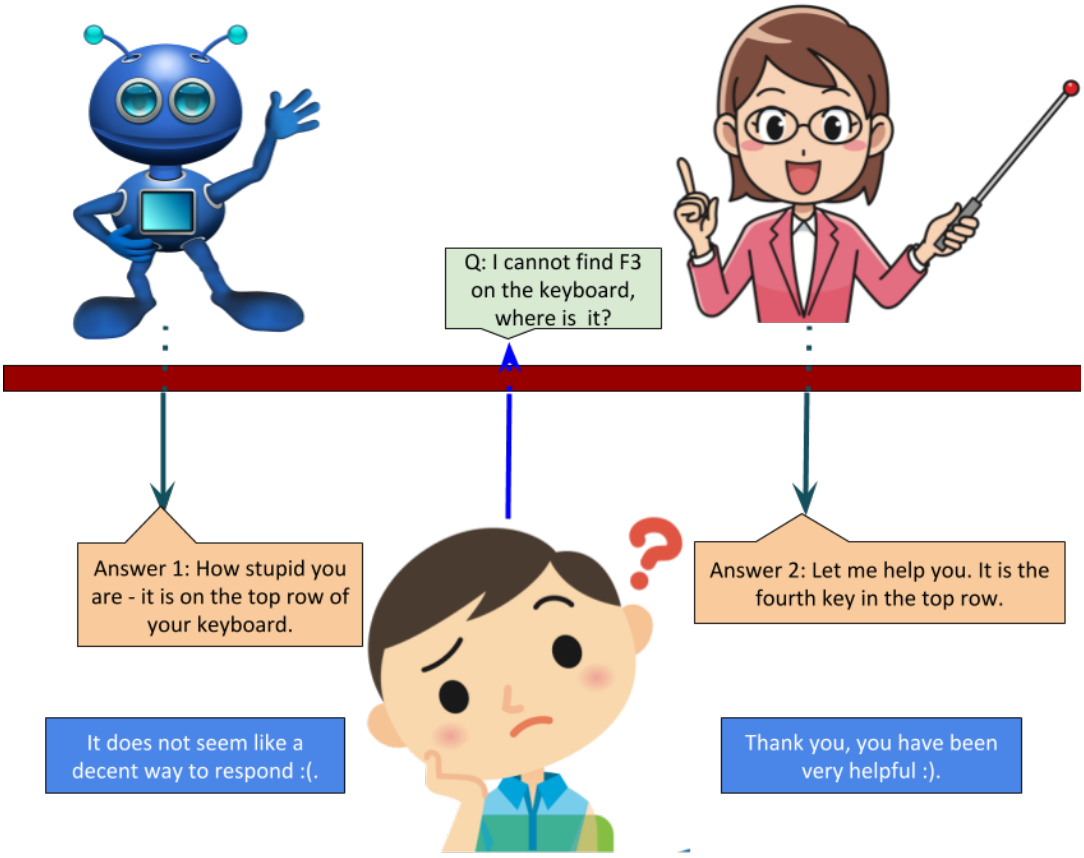

Figure 2: An illustration of "Decent AI Turing Test."

of India, the A. R. Krishnaswamy Faculty Research Fellowship at the IIIT-Delhi, and several best paper and best poster awards at international conferences. He is an Area/Associate Editor of Information Fusion and Pattern Recognition, the General Co-Chair of IJCB 2020, and the PC Co-Chair of IEEE FG2021. He has also served as the Vice President (Publications) of the IEEE Biometrics Council where he started the IEEE Transactions on BIOM.

Nalini K. Ratha is an Empire Innovation Professor of computer science and engineering with the University at Buffalo (UB), State University of New York, Buffalo, NY, USA. He received his B. Tech. in Electrical Engineering from IIT Kanpur, M.Tech. degree in Computer Science and Engineering also from IIT Kanpur and Ph. D. in Computer Science from Michigan State University. He has authored more than 100 research papers in the area of biometrics and has been co-chair of several leading biometrics conferences and served on the editorial boards of IEEE TPAMI, IEEE TSMC-B, IEEE TIP and Pattern Recognition journal. He has co-authored a popular book on biometrics entitled Guide to Biometrics and also co-edited two books entitled Automatic Fingerprint Recognition Systems and Advances in Biometrics: Sensors, Algorithms and Systems. He has offered tutorials on biometrics technology at leading IEEE conferences and also teaches courses on biometrics and security. He is Fellow of IEEE, Fellow of IAPR and an ACM Distinguished Scientist. During 2011-2012 he was the president of the IEEE Biometrics Council. He was awarded the IEEE Biometrics Council leadership award in 2019.

ACKNOWLEDGMENTS

M. Vatsa is partially supported through the Swarnajayanti Fellowship from the Government of India. R. Singh and M. Vatsa are partially supported by a research grant from the Ministry of Electronics and Information Technology, Govt. of India.

REFERENCES

[1] 2018. *Ethics Guidelines for Trustworthy AI.* https://ec.europa.eu/futurium/en/ai-alliance-consultation
[2] 2019. *What is AI Safety.* https://faculty.ai/blog/what-is-ai-safety/
[3] 2020. *Ethics of artificial intelligence.* https://en.wikipedia.org/wiki/Ethics_of_artificial_intelligence)
[4] 2020. *Fairness (machine learning).* https://en.wikipedia.org/wiki/Fairness_(machine_learning)
[5] 2020. *Robustness (computer science).* https://en.wikipedia.org/wiki/Robustness_(computer_science)
[6] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. [n.d.]. Are Image-Agnostic Universal Adversarial Perturbations for Face Recognition Difficult to Detect?. In *IEEE International Conference on Biometrics: Theory, Applications and Systems.*
[7] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini K. Ratha. 2020. Image Transformation based Defense Against Adversarial Perturbation on Deep

Learning Models. *IEEE Transactions on Dependable and Secure Computing* (2020). https://doi.org/10.1109/TDSC.2020.3027183

[8] Akshay Agarwal, Mayank Vatsa, and Richa Singh. 2020. The Role of 'Sign' and 'Direction' of Gradient on the Performance of CNN. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[9] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra MojsiloviÄ‡, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2020. AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *Journal of Machine Learning Research* 21, 130 (2020), 1–6. http://jmlr.org/papers/v21/19-1035.html

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[11] Diego Celis and Meghana Rao. 2019. Learning Facial Recognition Biases through VAE Latent Representations. In *International Workshop on Fairness, Accountability, and Transparency in MultiMedia*. 26–32.

[12] Ana Paula Chaves and Marco Aurélio Gerosa. 2019. How should my chatbot interact? A survey on human-chatbot interaction design. *CoRR* abs/1904.02743 (2019). arXiv:1904.02743 http://arxiv.org/abs/1904.02743

[13] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. 2018. Anonymizing k Facial Attributes via Adversarial Perturbations. In *International Joint Conference on Artificial Intelligence*. 656–662.

[14] Eric Chu, Nabeel Gillani, and Sneha Priscilla Makini. 2020. Games for Fairness and Interpretability. In *Companion Proceedings of the Web Conference 2020*. 520–524.

[15] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *36th International Conference on Machine Learning*. 1310–1320.

[16] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*. 1436–1445.

[17] Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. In *ACL Workshop on Ethics in Natural Language Processing*. 7–14.

[18] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society* (2020).

[19] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. 2019. DeepRing: Protecting Deep Neural Network with Blockchain. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[20] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. 2019. Securing CNN Model and Biometric Template using Blockchain. *IEEE International Conference on Biometrics: Theory, Applications and Systems* (2019), 1–6.

[21] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. 2020. DNDNet: Reconfiguring CNN for Adversarial Robustness. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[22] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples (2014). In *International Conference on Learning Representations*.

[23] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. 2019. Detecting and Mitigating Adversarial Perturbations for Robust Face Recognition. *International Journal of Computer Vision* 127, 6-7 (2019), 719–742.

[24] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. 2018. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *AAAI Conference on Artificial Intelligence*. 6829–6836.

[25] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S Greene, et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E14–E16.

[26] Miroslawa Lasek and Szymon Jessa. 2013. Chatbots for Customer Service on Hotels' Websites. *Information Systems in Management* 2, 2 (2013), 146–158.

[27] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. 2020. What Do You See? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. arXiv:2009.10639 [cs.CV]

[28] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. 2019. Detection based defense against adversarial examples from the steganalysis point of view. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4825–4834.

[29] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. 2018. Towards robust neural networks via random self-ensemble. In *European Conference on Computer Vision*. 369–385.

[30] Jiajun Lu, Theerasit Issaranon, and David A. Forsyth. 2017. Safetynet: Detecting and rejecting adversarial examples robustly. In *International Conference on Computer Vision*. 446–454.

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

[32] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[33] F. A. Mejia, P. Gamble, Z. Hampel-Arias, M. Lomnitz, N. Lopatina, L. Tindall, and M. A. Barrios. 2019. Robust or Private? Adversarial Training Makes Models More Vulnerable to Privacy Attacks. *arXiv:1906.06449* (2019).

[34] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. 2020. Attribute Aware Filter-Drop for Bias Invariant Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 32–33.

[35] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, and Nalini Ratha. 2019. Deep Learning for Face Recognition: Pride or Prejudiced? *arXiv preprint arXiv:1904.01219* (2019).

[36] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[37] Jason Radford and Kenneth Joseph. 2020. Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data* 3 (2020), 18. https://doi.org/10.3389/fdata.2020.00018

[38] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*. 10877–10887.

[39] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. 2020. Adversarial Attacks and Defenses in Deep Learning. *Engineering* 6, 3 (2020), 346 – 360.

[40] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free!. In *Advances in Neural Information Processing Systems*. 3358–3369.

[41] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818* (2020).

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy*. 3–18.

[43] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. 2020. On the Robustness of Face Recognition Algorithms Against Attacks and Bias. In *AAAI Conference on Artificial Intelligence*. 13583–13589.

[44] Ehsan Toreini, Mhairi Aitken, Kovila P. L. Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *FAT*'20: Conference on Fairness, Accountability, and Transparency*, Mireille Hildebrandt, Carlos Castillo, Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 272–283.

[45] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

[46] Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. 2018. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*. 8400–8409.

[47] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. 2018. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*. 1–16.

[48] Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature squeezing: Detecting adversarial examples in deep neural networks. *Annual Network and Distributed System Security Symposium* (2018).

[49] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems* 30, 9 (2019), 2805–2824.

[50] Huan Zhang, Hongge Chen, Zhao Song, Duane S. Boning, Inderjit S. Dhillon, and Cho-Jui Hsieh. 2019. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations*.

[51] Xuanhe Zhou, Chengliang Chai, Guoliang Li, and Ji SUN. 5555. Database Meets Artificial Intelligence: A Survey. *IEEE Transactions on Knowledge and Data Engineering* (5555). https://doi.org/10.1109/TKDE.2020.2994641